Big data *para científicos sociales. Una introducción*

José Manuel Robles, J. Tinguaro Rodríguez, Rafael Caballero y Daniel Gómez
(Madrid, Centro de Investigaciones Sociológicas, 2020. Cuadernos Metodológicos, 60)

Este libro se inscribe en la colección Cuadernos Metodológicos elaborada por el Centro de Investigaciones Sociológicas (CIS), galardonada por la Unión de Editoriales Universitarias Españolas (UNE) con el Premio Nacional de Edición Universitaria de 2009 por ser considerada por el jurado «la colección más importante de metodología sociológica en el mundo hispánico». En concreto este libro pretende realizar una introducción a lo que denomina *big data* para todo aquel investigador social que pueda estar interesado en aproximarse a esta materia y su gran potencial para incorporarla a sus trabajos e investigaciones.

Tras las innumerables revoluciones tecnológicas que han tenido lugar durante las últimas décadas, una de las más recientes y con mayor impacto está siendo la revolución de los datos. Debido a la alta presencia de Internet en nuestras vidas, dada la aparición de los *smartphones*, de las redes sociales, del Internet de las cosas, etc., nos encontramos con una inmensurable cantidad de información de cuyo valor se está tomando consciencia progresivamente. En el momento actual son innumerables las empresas que hacen uso de todos estos datos para sus estudios de mercado, para sus procesos industriales o para ofrecer distintas funcionalidades a sus clientes. Pero estos datos también resultan de enorme utilidad para la investigación científica, y en concreto, para las ciencias sociales, cuyo enfoque dominará en esta presentación de las técnicas de *big data*. Mediante esta introducción se pretende dar pie a la reflexión sobre el desafío metodológico y epistemológico que supone el *big data*, a la vez que se da a conocer al científico social las posibilidades y técnicas existentes para el análisis de datos de cara a que pueda participar en equipos interdisciplinares para abordar las ciencias sociales mediante soluciones *big data*. Se entiende que para un científico social lograr la profundidad de conocimiento técnico de un matemático o un informático es una tarea muy complicada, pero comprender las bases del trabajo que realizan en lo que se refiere al análisis de datos puede ser muy útil para unir sus fuerzas y conocimientos para aproximarse e interpretar la realidad social.

Este libro se estructura en cuatro capítulos, comenzado por una presentación de qué es el *big data* y su encaje en la investigación de ciencias sociales. En esta se narra la evolución histórica del almacenamiento de los datos, desde las primeras bases de datos rudimentarias hasta el modelo relacional SQL y sus problemas de escalabilidad vertical, que dada la emergente cantidad de datos existentes ha dado lugar al *big data* y a su capacidad de escalabilidad horizontal que planteó Google por primera vez. Así, cuando hablamos de *big data*, hablamos de grandes cantidades de datos cuyas características principales son

lo que se da a conocer como las tres uves: volumen, velocidad y variedad, a las que habría que añadir veracidad y valor. El *big data*, por tanto, nos debe permitir trabajar con grandes cantidades de datos, que se generan y con los que hay que trabajar a cierta velocidad, que pueden encontrarse en diferentes formatos y de los cuales se busca extraer una información veraz y que nos proporcione valor. Esta tarea se inscribirá en la ciencia de datos, disciplina constituida por un conjunto de técnicas estadísticas y de inteligencia artificial para extraer información de los datos, basadas en las matemáticas, la informática y, en este caso, las ciencias sociales como ámbito de aplicación. A lo largo de este libro se aportarán los conocimientos de las dos primeras disciplinas de cara a que el científico social pueda participar en estos equipos interdisciplinares de ciencias de datos con el objetivo de buscar regularidades en las poblaciones para tratar de comprender la acción social, aunque su explicación puede ser compleja por medio de las técnicas estadísticas. Me parece importante destacar que por tratarse la ciencia de datos de algo bastante novedoso y que lleva pocos años popularizándose, y más en lo relativo a las ciencias sociales, sus posibilidades en este campo aún están por descubrir. Es posible que en la actualidad estas técnicas se estén centrando más en ámbitos descriptivos que explicativos, pero hay que tener en cuenta que es una disciplina con un enorme potencial y un largo camino por recorrer. Las tareas de análisis que ahora parecen sencillas y comunes antes podían resultar muy complejas o marginales, por lo que creo imprescindible trabajar en este campo con la esperanza de que sea una herramienta aún más útil y potente para la investigación social.

En el segundo capítulo ya comienza a adentrarse en materia más técnica, dando a conocer las diferentes fuentes de datos existentes con que se puede trabajar, las cuales son las redes sociales, los datos de páginas webs y los ficheros disponibles para descarga en Internet. También nos explica los diferentes formatos en que se encuentra la información en estas fuentes y comienza a adentrarse en el mundo de la programación para acceder a estos datos mediante el api de Twitter o *web scrapping*. Para ello se centrará en el lenguaje Python a través de Jupyter Notebooks, dando también la opción de ver estos códigos en R en un repositorio GitHub de libre acceso. Pese a no ser el objetivo del libro, que trata de ser una introducción breve y concreta sobre la materia, echo en falta una breve aproximación para el lector que pueda ser más ajeno a esta sobre la tarea de programación. Explicar que es una tarea delicada, que se basa en acudir constantemente a Internet para ver cómo plasmar los procesos que queremos realizar, y en la cual los errores de código son lo más común del mundo y su resolución supone una parte importante del trabajo, creo que facilitaría al lector empatizar con esta labor. Igualmente, las secciones de código planteadas a lo largo del libro están bastante bien explicadas paso a paso y de una manera muy ordenada, facilitando la relación entre el lector y la programación para futuras tareas que pudieran implicar el conocimiento de esta. En algunas ocasiones, la forma de algoritmos muy concretos puede hacerlos difíciles de seguir y, aunque esto no es necesario para esa aproximación, me parecería útil un glosario de funciones más concretas que se empleen en el libro para poder tratar de comprenderlos. Aun así, no es una tarea difícil para el lector, e incluso puede ser más ligera, acudir a manuales digitales o a Internet para esto, ya que no es exigible para este tipo de libro que realice las funciones más propias de un manual.

El tercer capítulo versa sobre el almacenamiento de la información, dando a conocer las posibilidades de la nube o *cloud* y el funcionamiento de las bases de datos relacionales SQL y no relacionales, centrándose en MongoDB. Se muestra cómo crear las bases de datos, cómo añadir y borrar registros, cómo extraer la información de estas, cómo visuali-

zarla… de cara a que el científico social conozca la estructura en que se almacenan los datos y cómo se trabaja con ellos, un aspecto crucial y que queda cubierto casi por completo en este capítulo. Además, trata los *softwares* probablemente más extendidos y útiles para este aspecto, que podemos encontrar en casi cualquier registro de datos existente.

Por último, y tal vez el capítulo más interesante, encontramos una presentación del tratamiento y análisis computacional y sus diversas técnicas para la extracción de conocimiento útil y creación de valor añadido. Estas técnicas se basan en herramientas matemáticas y/o computacionales para la creación de modelos explicativos y/o predictivos. Se trata de técnicas escalables horizontalmente que nos permiten modelizar la relación de dependencia entre unas determinadas variables independientes y dependientes de naturaleza numérica o categórica. Aparecen divididas en técnicas de *machine learning* y análisis de redes sociales.

En lo que respecta a las primeras se muestran como programas informáticos creados para resolver problemas mediante la generación de algoritmos de clasificación o predicción a partir de los datos, buscando la optimización de determinadas medidas de eficiencia asociadas al programa. Una definición muy acertada y comprensible que captura totalmente la esencia de estas técnicas. Además, comenta la diferencia entre aprendizaje supervisado y no supervisado, así como sus pros y contras. También define las medidas de eficiencia más comunes y la importancia del rendimiento de estos algoritmos para que puedan ser soportados por las computadoras en tiempos normales. Del mismo modo trata aspectos más específicos, pero de crucial importancia si se quiere saber cómo se trabaja con estos programas, como el peso de los datos de entrenamiento en la definición del algoritmo y la importancia de comprobar la eficiencia sobre un conjunto de test, técnicas para lidiar con este problema como la validación cruzada, los problemas de clasificación no equilibrada o los peligros del sobreajuste. Se trata en todo momento de dar a conocer los aspectos que se tienen en cuenta en un análisis de datos para poder seguir procesos de este tipo. A continuación, se centra en determinadas técnicas, en concreto: el algoritmo de los k vecinos más cercanos, los árboles de decisión, el clasificador bayesiano, las redes neuronales, las máquinas de soporte vectorial, *random forest* y el algoritmo de k medias para *clustering*. En todos estos ejemplos se explica el funcionamiento de la técnica de manera básica, pero incluso mostrando las ecuaciones matemáticas en que se basa y que explican su funcionamiento, acompañado esto de un ejemplo práctico en que se aplica la técnica a un problema concreto para ilustrar su funcionamiento y los resultados que puede ofrecer. Respecto a esto me parece importante tratar de quitarle hierro en el texto a la comprensión completa de los algoritmos, formulas y funcionamiento, que debe fundamentarse principalmente en unas nociones más generales. Un conocimiento pleno de todo lo expuesto en una sola lectura puede ser muy complicado, pero desde luego esta exhaustividad es muy útil si el autor desea trabajar específicamente con algunas de estas técnicas. Del mismo modo creo que sería conveniente para algunas de estas técnicas una mayor presencia de ilustraciones sobre ejemplos sencillos que permitan aprehender el funcionamiento básico.

En lo que respecta al análisis de redes sociales se hace una presentación de estas técnicas y su conveniencia para muchas tareas, en las que se puede obtener una mayor información de la relación entre las unidades de información que de la información intrínseca a cada una de estas. Esto se sigue de la definición de sus conceptos básicos, las diversas formas de representación y de las diferentes medidas y formas de análisis con que pode-

mos abordar estos problemas para extraer información útil y valiosa, dando especial importancia a las medidas de centralidad, análisis topológico y de comunidades, que también ilustran mediante ejemplos, especificando la información que podemos extraer mediante cada una de ellas. Desde luego, la inclusión del análisis de redes sociales en el libro me parece muy acertada, dado el alto potencial que tiene para la investigación de ciencias sociales y la gran capacidad para abstraer información, sobre todo si se combinan con técnicas complejas de recogida de datos de redes sociales o del Internet de las cosas.

La elaboración de este libro me parece una muy buena iniciativa, con un planteamiento muy acertado para tratar de involucrar entre sí las diferentes disciplinas que participan bajo el marco del análisis de datos para ciencias sociales. La estructuración del libro lo hace muy sencillo de seguir y fácilmente digerible para cualquier persona, indiferentemente de su grado de conocimiento en esta materia. Igualmente creo relevante hacer hincapié dentro del libro para hacer saber al lector que no debe frustrarse por la difícil comprensión de algunas secciones más matemáticas o de programación que pueden asustar a una persona alejada de la materia, ya que esto no condiciona su capacidad para participar de proyectos en que se trabaje con estas técnicas y sobre los que podrán aportar sin lugar a dudas sus conocimientos sobre otros ámbitos si comprenden los conceptos básicos que se presentan en este libro. El contenido del libro pasa por prácticamente todas las raíces de esta disciplina, que se trata de poner en valor dando un marco general muy adecuado para la participación en proyectos de análisis de datos sociales por parte de los científicos sociales. Tal vez, lo que echo más en falta es un capítulo o epígrafe dedicado al preanálisis de los datos, muy importante para el trabajo con estas técnicas, dada la necesidad de conocer los datos con que se va a trabajar para su posterior tratamiento. Del mismo modo, creo importante dar a considerar que, aunque los datos son algo bastante neutro, a la hora de generar los algoritmos, el etiquetado de datos o la búsqueda e interpretación de los resultados, pueden plasmarse prejuicios o consideraciones propias del autor que pueden alterar la información real que contienen los datos, la cual suele llevar a conclusiones muy concretas sobre los datos específicos con que se trabaja.

En resumen, se trata de un libro muy interesante, ameno y completo que invita a la reflexión del lector sobre su posible papel en un nicho bastante concreto de estudio, pero que, poco a poco, se expandirá e ira aumentando de peso, como es el análisis de datos para ciencias sociales. Aboga por la participación de los científicos sociales en proyectos interdisciplinares que utilicen estas técnicas y en los que necesitarán de los conocimientos básicos sobre ellas para poder exprimir al máximo los datos existentes y obtener el mayor conocimiento posible sobre la realidad social. La cuestión sería: ¿estamos cerca de estos límites en el provecho que se le puede sacar a la información mediante el análisis de datos? O, por el contrario, ¿son los estudios e investigaciones que se llevan a cabo actualmente solo la punta del iceberg de lo que podrían ser una gran gama de informaciones y explicaciones sociales que podemos obtener por medio de este tipo de técnicas? Es seguro que el tiempo nos lo dirá, siempre que el desempeño en esta disciplina de ciencia de datos para ciencias sociales siga creciendo.

por Alejandro ECHÁNIZ-JIMÉNEZ
Universidad Complutense de Madrid
aechaniz@ucm.es